# Chapter Three

**Methodology**

## Introduction

This chapter discusses the research methodology used in the development of a proposed hybridized geometric similarity measure with k-NN model for improved case retrieval in CBR systems. The materials and procedures involved in achieving this model are also discussed.

In order to further improve the efficiency of case retrieval in CBR systems, this research aim to create a hybrid similarity measure by combining the Euclidean distance measure with other proven geometric similarity measures, such as Cosine similarity and Jaccard similarity, to produce a more accurate and comprehensive measure of similarity between cases. The hybridization of the geometric similarity measures is intended to integrate the benefits of established similarity measures while mitigating their drawbacks. A step-by-step approach to achieve this is discussed. Starting with the data collection approach used to support the testing of the CBR system, the mathematical equations used for the setup/development of the hybridized similarity measure leading to the improved case retrieval in CBR systems to the algorithmic setup for each stage of the design.

## Research Approach

The two main approaches to research are: qualitative and quantitative. The two approaches are slightly different from each other.

* + - 1. **Quantitative research**: This is majorly based on the collection and analysis of numerical data. It is often used to test hypotheses and to generalize findings to a larger population. Quantitative research methods include experiments, surveys, and statistical analysis.
      2. **Qualitative Research**: This research approach is based on the collection and analysis of non-numerical data, such as interviews, observations, and documents. It is often used to explore complex phenomena and to gain insights into people's experiences and perspectives. Qualitative research methods include ethnography, case studies, and thematic analysis.

This research falls under the qualitative approach.

The specific activities under this approach involves the following steps:

* + - 1. Development of a dataset of cases that can be used to evaluate the performance of the hybridized geometric similarity measures. The dataset is a representative of the real-world domain in which the CBR systems will be used.
      2. Design of a new hybridized geometric similarity measure. The new measure is based on a combination of existing geometric similarity measures, such as the Euclidean distance measure, Cosine similarity measure and Jaccard similarity measures.
      3. Implementation of the hybridized geometric similarity measure in a Python programming language.
      4. Evaluation of the performance of the hybridized geometric similarity measure on the developed dataset by conducting experiments. The experiments compare the new measure to standard similarity measures.
      5. Furthermore, collection of data on the computational cost and explainability of the different similarity measures.
      6. Finally, the results of the experiments are analyzed so as to identify if the hybridized geometric similarity measure for case retrieval achieve the set research objectives.

A replication of this study can be done by; downloading the dataset from a public repository, cloning of the code repository, installation of the required dependencies, running of the experiments to evaluate the performance of the hybridized geometric similarity measures on the dataset, finally, analysis of the results of the experiments and compare the performance of the different similarity measures. These steps is to ensure the study can be replicated and results verified by the research community.

## System Design

Over time, different concepts and techniques have been incorporated into CBR to record and index cases and then search the cases in the case base in order to identify cases that might be a solution to newly presented cases. Depending on the kind of decision to be made, designing any of the steps of a CBR system involves addressing different issues. Designing an efficient case retrieval method in a CBR system involves activities such as the representation of cases, search the case base to retrieve cases, match retrieved cases by using similarity metrics, even ranking of cases, and adaptation.

### Case Representation

Case representation is a crucial aspect of a CBR systems as it determines how cases (knowledge) are stored and processed. In fact, this is the first activity that needs to be performed. It is actualized by capturing the essential information about a case in a structured and logical manner. Both base case and new cases needs to be represented.

Hence, in this research, while modelling the case library, base cases which consists of a set of problem-solution features structure were firstly established and stored in the case base/library.

The base case model is represented as:

Case = {*Problem Description*, *Solution*}

* + - * 1. **Problem Description**: This represents a detailed description of the problem that the case addresses. It is represented by the problem feature subsets and consists of a number of principal problem features. The problem description is described by a sequence of problem features, (*f1*, *f2*, *f3*, . . . *fn*).

The problem features (*fi*) of the problem description are represented as attribute-value pairs:

*fi* = (*ai, vi*)

Where *ai* is the attribute of the problem which are defined in the problem features vocabulary and *vi* are the values related to each attribute. As it is the case in this research, in most instances attributes are determined by domain experts.

The values may involve numerical values, categorical labels or in some instances a combination of both. This include the context, symptoms and other relevant observations.

* + - * 1. **Solution**: The solution represents the result of the problem description. It is the solution or set of solutions that were implemented to address the problem. This include specific actions, interventions or strategies.

Note that, while the structure of the base case consists of the problem description and the solution, the structure of the new case only consists of the problem description. Figure 3. 1 represents the structure of the base case. The problem description is represented by a subset of problem features while each of the problem feature is represented by a principal attribute and a corresponding value. Hence, depicting an attribute-value pair.

Although, there are several case representation methods, such as relational database case representation, attribute-value representation, predicate-based representation and even soft computing case representation methods. Each of them has its own set of advantages and disadvantages. The attribute-value representation approach has been discovered to be appropriate for this study**.** Using attributes with characteristic values makes it straightforward to represent the instances gathered. The proposed hybridized geometric similarity measure utilizes the attribute-value representation of cases to compute similarity. By considering the primary range of features, the hybridized measure can capture more nuanced relationships between cases and improve the accuracy of case retrieval. The use of attribute-value representation to express cases is justified since it makes it easier to specify cases and allows for the use of the k-NN retrieval method.

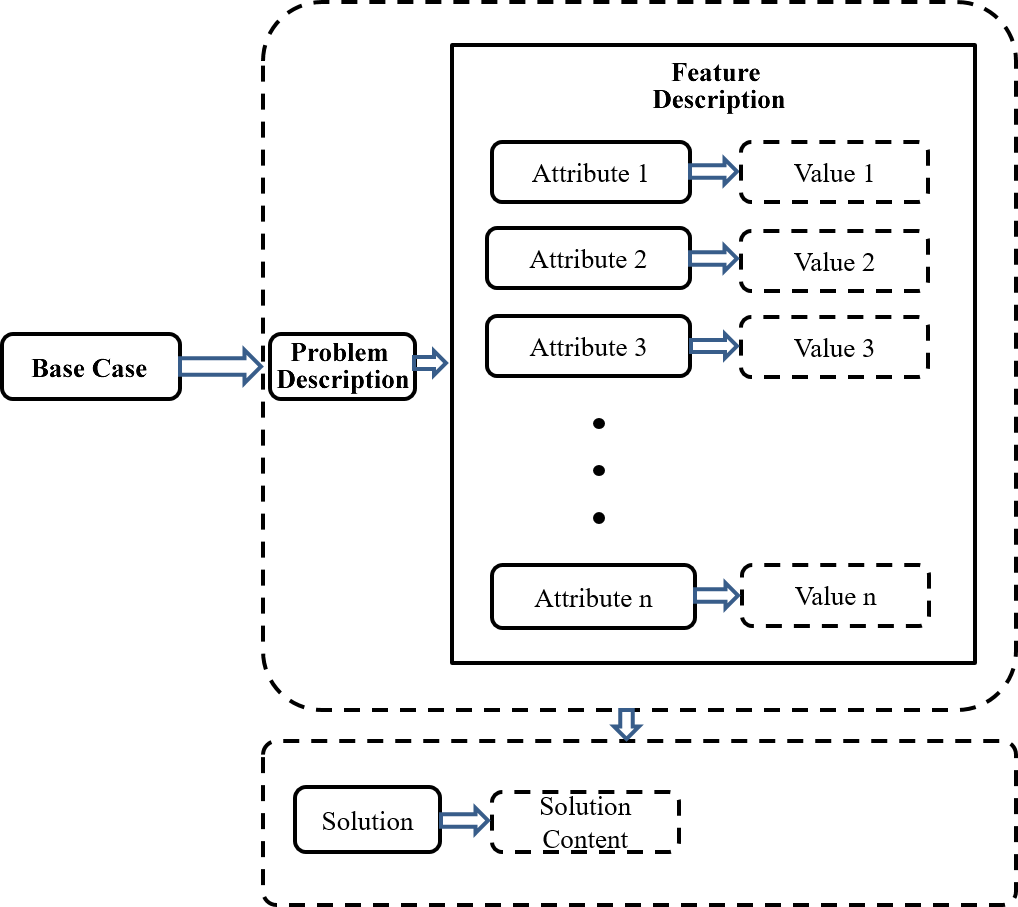


Fig 3. 1 Structure of a Base Case (Source: Researcher, Omonijo O. 2024)

The case representation features for the system are as follows:

1. **Case ID**: Each case has a unique case number. The number uniquely identifies a case or a problem. It acts as a primary key for accessing and organizing cases within the case base.
2. **Case Feature**: Each case or disease has a series of descriptions (symptoms) which are the attributes of the problem and are usually more accurate, specific and easily understood with relevant features. It serves as the primary input for similarity measures, enabling the system to identify similar cases based on their descriptions.
3. **Case Alias**: Although, this is optional, the alias attribute contains a list of all abbreviations and non-keywords. This enables the system to still perform in a case where user uses aliases or abbreviations in the case query. Provides additional keywords, synonyms, or abbreviations associated with the case, enhancing search capabilities. The aliases are parsed through the description so as to match the brief description with the cases in the case-base. Allows users to find cases using familiar terms or jargon, improving accessibility.
4. **Solution**: This is the action or solution to the problem. It contains the recommended solution, resolution, or outcome for the problem described in the case.

### The System’s Case Base

CBR can only be used if there is a case base of relevant cases to draw from. As such, the case base needs to be structured in such a way that cases can be easily retrieved. Each case contains a set of description of the problem and the associated solution. The problem description subset of the case is represented by the problem feature subsets and a number of principal attributes are associated to each feature. The case base is deployed using a Comma-Separated Value (CSV) file. The structure of the case base is represented in Table 3. 1.

The case base is structured such that;

* + - 1. Each case is represented in a structured format that captures all the relevant information and/or attributes about the case. The information includes the Case ID, Case Feature, Case Alias (which is optional) and the Solution.
      2. The cases are indexed using the Case ID in a way that it makes it easy for the model to find the most similar cases. This involves the creation of unique identification which is used to uniquely refer to a case.
      3. Cases are organized to make it easy for the system to understand the relationships between the cases. This was done by organizing the Case Feature by disease type and solution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **FIELD NAME** | | **DATA TYPE** | **REMARK** |
| Problem Description | Case ID | | Varchar | Unique |
| Case Feature | Case Attributes | Text |  |
| Attribute Value | Boolean |  |
| Case Alias | | Text |  |
| Solution | Solution | | Text |  |

Table 3. 1 Case Base Structure

### Model of the Proposed Hybridized Similarity Measure

In CBR, similarity is a bridge between case representation and retrieval30. It measures the similarity amongst pairs of problem descriptions of cases, *Sim*(*pm*, *pq*), such that, the solutions of the base cases *Cm* can be used to find the solution of the query case *Cq*where, *pm* represents the problem description of cases in the case base and *pq* is the problem description of the query case.

The method employed in this research involves the calculation of the local and global similarity values between the base case and the query case. The highest ranked solution of the base case whose similarity value is above the threshold is considered as the solution to the case. In the event that the similarity value does not surpass the threshold, an expert will be required to update the solution and store in the case base for future episodes.

The model which makes up the proposed hybridized similarity measure for case retrieval in CBR system is as follow.

* 1. **Measurement of Similarity**

When a new problem is entered into the system, similar cases are sorted out by measuring the features of the problem description against the features of the problem description of cases in the system’s base case. Values obtained from similarity measurement will determine the degree of similarities between the new problem and the base cases on a case-by-case basis by comparing the features in the new case with those in the base cases. To carry out the comparisim, a hybridized similarity measure was used.

The similarity was computed using the local-global principle in two stages: first, at the feature level (feature matching), which required taking into account the values of common attributes, and second, at the case level, which required combining the results according to the case representation.

* + - * 1. **Calculation of the Individual Similarity Scores (Local Similarity)**

Distance measurements are often used to compute the local similarity functions. The local similarity function employs an attributes to attributes comparisim between a new problem and the case stored in the case base.

To model the local similarity function of the proposed hybrid similarity measure, common geometric similarity measures such as Euclidean distance, Cosine and Jaccard similarity measures were hybridized. This was used to compute the geometric similarity score between the attributes, “*aj*” of the new case’s problem description, “*xk*” and that of the case in the case base, “*yk*”.

A local similarity measure *simj* was then defined between the attribute values *xk*(*aj*) and *yk*(*aj*):

*simj*(*xk*(*aj*), *yk*(*aj*)) ≡ *sim*(*x, y*)

This is the result of the feature matching.

The geometric similarity measures of Euclidean distance, Cosine and Jaccard similarity measures are represented in equation 3. 1, 3.2 and 3. 3 respectively.

Euclidean = ***Dist*(*x, y*)=**  (3.1)

Where; *Dist( )* is the distance that exist between the two compared vectors.

*x* and *y* represents the attribute values of the feature being compared.

Cosine = ***Cos(θ)* = *Cos(x, y)* =**  (3.2)

Where;

***.*** represents the dot product of the vectors.

*x* and *y* represents the attribute values of the feature being compared.

***|| x ||*** and ***|| y ||*** represents the lengths of *x* and *y* respectively.

θ is the angle between the vectors

Thus, ***|| x ||***and ***|| y ||***represent the Euclidean form of vector *x* and *y*

Jaccard = ***JS(x, y)*** = (3.3)

Where;

*x* and *y* are the sets of features being compared.

|***x******y***| represents the number of attributes that are shared between sets *x* and y (the intersection).

|***x*** ∪ ***y***| is the total number of unique attributes in both sets *x* and *y* (the union).

| ***…*** | represents the cardinality of the set of attributes.

Equation 3.4 is used to achieve the result of the feature matching which is referred here as the hybridized local similarity function score.

***hsim*(*x, y*)** = (3.4)

where:

***hsim*(*x, y*)**is the hybridized local similarity score.

is the Euclidean distance which measures the value of the feature between attributes *x* and *y*.

is the Cosine angle which measures the value of the feature between attributes *x* and *y*.

is the Jaccard similarity measure which measures the value of the feature between attributes *x* and *y*.

Although, Equation 3.4 is a hybrid score, it can only be used to measure one attribute.

* 1. **Summation (Generalization) of the Individual Similarity Score (Global Similarity)**

Combining the local measures at the global level requires scaling up the distance measures from the single dimensional approach which only entails dealing with one attribute, ***hsim*(*x****(aj)****, y****(aj)***)** ≡ ***hsim*(*x, y*)** to the multidimensional approach which combines all the case features, ***HSim*(*X, Y*)**.

To achieve this, the features of each case in the case base was matched against the features of the queried case.

Hence, the geometric similarity measures of Euclidean distance, Cosine and Jaccard similarity measures as represented in equation 3. 1, 3.2 and 3. 3 respectively will become:

Euclidean = ***Dist*(*X, Y*)=**  (3.5)

Where;

*Dist( )* is the distance that exist between the two compared vectors.

*i* is the attribute for each case.

*n* is the number of attributes for each case.

*xi*, *yi* represent the *i*th attribute in the vector *x* and *y* respectively.

Cosine = ***Cos(θ)* = *Cos(X, Y)* =**  (3.6)

Where;

***x . y* =**

***|| x ||***is length of the vector *x*and represented as ***|| x || =***

While ***|| y ||***is the length of the vector ***y*** and represented by ***|| y || =***

θ is the angle between the vectors

*x* represents values from base case where *xi = x1, x2, x3, … xn*

*y* represents values from target case where *yi = y1, y2, y3, … yn*

*n* is the number of features in each vector.

*xi* and *yi* are the *i*-th attribute values of the feature in vectors *x* and *y*, respectively.

Thus, ***|| x ||***and ***|| y ||***represent the Euclidean form of vector *x* and *y*

Jaccard = ***JS(X, Y)*** = (3.7)

Where;

|***x******y***| represents the number of attributes that are shared between sets *X* and *Y* (the intersection).

|***x*** ∪ ***y***| is the total number of unique attributes in both sets *X* and *Y* (the union).

is the sum of values that are present in both sets (the intersection).

The Jaccard Similarity, like the traditional Jaccard Similarity, focuses on the proportion of shared items while considering their relative importance.

From equation 3.5, 3.6 and 3.7, the hybrid similarity measure is thus given as:

***HSim(X, Y)*** = (3.8)

Equation 3.8 gives a global representation of the equation. But then, because each similarity measure in the hybrid similarity measure, ***HSim(X, Y)*** have different degree of contribution to the similarity measure, weights are assigned to each of the similarity measures in equations 3. 5, 3. 6 and 3. 7 based on the perceived importance or relevance of the similarity measure to the domain. This will ensure that a balance exist amongst the three measures such that the effect of the orientation magnitude distance (Euclidean distance), (Cosine similarity) and commonalities (Jaccard similarity) between new case and base cases are considered. To ensure normalization, the total sum of the values of the three weights are set to be equal to one (1). This is to ensure that the value of the hybridized similarity measure does not exceed the normalized values of 0 and 1.

Primarily, there are two types of weights allocation methods: subjective allocated weights (delphi method, statistical method, hybrid weight mean method) and objective allocated weights (entropy method, neural networks algorithms, genetic algorithms). While the latter is influenced by the weakness of subjective weight, the former is influence by decision-makers.

In this research, weights are applied subjectively. Weights are assigned to each similarity measure in equations 3.5, 3.6 and 3.7 based on the perceived importance or relevance of the weights to the similarity measure.

With respect to equation (3.5), the formula for the Weighted Euclidean Similarity function is given as:

***WED(X, Y) = 1 -***  (3.9)

Where;

is the weight assigned to the k-th feature, indicating its relative importance.

From equation (3.6), the formula for the Weighted Cosine Similarity function is given as:

***WCS(X, Y )=***  (3.10)

Where;

is the weight assigned to the *i*-th feature, indicating its relative importance.

Weighted Cosine Similarity (WCS) captures the similarity in the direction of vectors, considering both feature values and their weights.

While the formula for the Weighted Jaccard Similarity function with respect to equation (3.7) is given as:

***WJS(X, Y) =***  (3.11)

Where;

is the weight assigned to the *i*-th feature, indicating its relative importance.

The weighted average method was used to integrate the individual weighted similarity scores for each case form equation 3.9, 3.10 and 3.11. This hybrid similarity score reflects the combined similarity based on the weighted features.

***WHSim(X, Y)*** = (3.12)

Where:

***WHSim*(*X, Y*)**is the hybrid similarity measure

is the weighted Euclidean measure between cases X and Y for the *i*-th case.

is the weighted Cosine measure between cases X and Y for the *i*-th case.

is the weighted Jaccard similarity measure between cases X and Y for the *i*-th case.

Note: in all cases of the similarity measures, the values of *wi* can be tuned to achieve the best performance on a particular dataset.

The hybrid similarity scores directly represent the overall similarity measure based on the weighted features and it is normalized to values between 0 and 1.

* 1. **Normalization of Retrieved Cases**

The use of normalized values eases the desire for comparisim, most especially at the global level. As a result, the numerical attribute values were normalized before use in the local similarity measure.

The composite combination of the local results produced similarity measures which are normalized in [0, 1] and are comparable. The results of all the features were averaged since the idea is to have global result values in [0, 1] for the final outcome of *WHSim(X, Y)*.

The similarity, *WHSim(X, Y)* is expressed as a value between 0 and 1 by using the following equation.

***WHSim(X, Y)*** = (3.13)

Hence, the similarity is defined as a value in the range [0,1], where 1 represents the maximum similarity (*WHSim(X, X)* = 1) and 0, the least.

* 1. **Rank the Retrieved Cases**

Tasks (1) and (2) will return set of cases with different similarity degrees (values) to the query case. As a result, cases are ranked according to their preference relation which was induced by their utility for solving the case. This creates a ranked list where the most similar cases to the query case appears at the top. Hence, all retrieved case sets were sorted in descending order of their *WHSim* scores. Case whose similarity values is higher than a set benchmark (threshold, τ) value will be considered for the next task.

* 1. **Retrieve Cases Using k-NN**

Results from the ranking tasks in (3) generates a descending ordered list, *RC* of cases contained in the memory, *RC* = [] where cases which exceeds the set threshold are considered. This implies that, the list containing all of the cases in memory, where *n* = *k*, or a small number of *n* whose value is greater than the similarity degree (threshold) are considered as potential solutions. Since the ranking is done in a decreasing order then, the top *k* which is the closest cases to the new case using the algorithm suggested poses as the solution to the queried case.

In this research, the cases which are above the threshold value are considered to give a solution. The top k cases which exceeds the threshold are selected from the ranked list to form the k-nearest neighbours. This implies that a value for “*k*” which represents the desired number of nearest neighbours to be retrieved is chosen. But then, instances where no case exceeds the threshold, the topmost cases are considered by an expert and appropriate solution adapted and saved in the case base in case of future episodes.

To find the k-NN:

* + - * 1. Define a value for the threshold, τ.
        2. Define a value for “*k*”, the desired number of nearest neigbhours.
        3. Sort the cases in descending order of their similarity scores (SS).

SSi = S(xi, q) for each xi in X

* + - * 1. Select the top “k” cases from the sorted list as the k-NN:

SS(q) = {xi | i ∈ argsort (SSi) [0 : k]}.

The hybridization occurred locally in equation 3.4. It was globalized in equation 3.8 but was weighted in equation 3.12 and normalized in equation 3.13, thereby giving a multiple weighted geometric measure which combined traditional similarity measures to create a more comprehensive similarity assessment.

The weights assigned to each measure can be adjusted based on the domain and data characteristics to prioritize different aspects of similarity.

From the foregoing, the system uses a CBR approach embedded with the proposed hybridized geometric similarity measurement method to determine how similar existing and new cases are. The user enters the problem description of the new problem into the system, this contains the features of the problem. The system then uses the proposed hybridized similarity measure which uses the k-NN method to determine the local-global similarity values between the new case data (user input) and the old cases in the case base. The case which have the highest similarity value (which should be more than the threshold) will be utilized as the answer to the new case. If the computed similarity score does not surpass the set threshold value then, the case solution will be updated by an expert and will be stored in the case base in case of future episode. The system outputs the solution of the query entered using the proposed algorithm.

### Conceptual Model of the Proposed System

The basic idea in the retrieval of cases in a CBR system is to ensure that the accurate case(s) are retrieved when using a similarity measure and efficiently too. Thus, the proposed CBR system aim to utilize a case base of previously encountered cases and a hybridized geometric similarity measure to retrieve the most similar cases based on new case characteristics, features or attributes.

To achieve this, the following core components were employed by the system:

**Proper Case Representation**: Representation involves how cases are stored and case-base constructed. Thus, the capture of the essential information about a case must be in a structured and meaningful way. Cases and their attributes are usually domain specific hence, case representation differs from domain to domain but may have few features in common. Proper representation of cases in most instances requires the knowledge of a domain expert. Generally, the appropriate representation method is selected according to the different application fields1.

**Case Base**: This is a repository of well-structured cases containing information about the data set to be used. Data to be stored in a case base can have as many features as needed. In this study, cases to be stored in the case base are stored as a pair of problem description and corresponding solution. The problem description subset of the case is represented by the problem feature subsets and made up of a number of principal attributes associated to each feature.

**Hybridized Geometric Similarity Measure**: A novel similarity measure that combines weighted traditional geometric similarity measures, such as Euclidean distance, Cosine similarity, and Jaccard similarity, with the aim of complementing the strength of each of the similarity measure with the weakness of the other so as to have a more efficient case retrieval model that capture the complex relationships between the cases was developed.

**Similarity Ranking**: The retrieved cases are ranked in a decreasing order based on their global similarity scores.

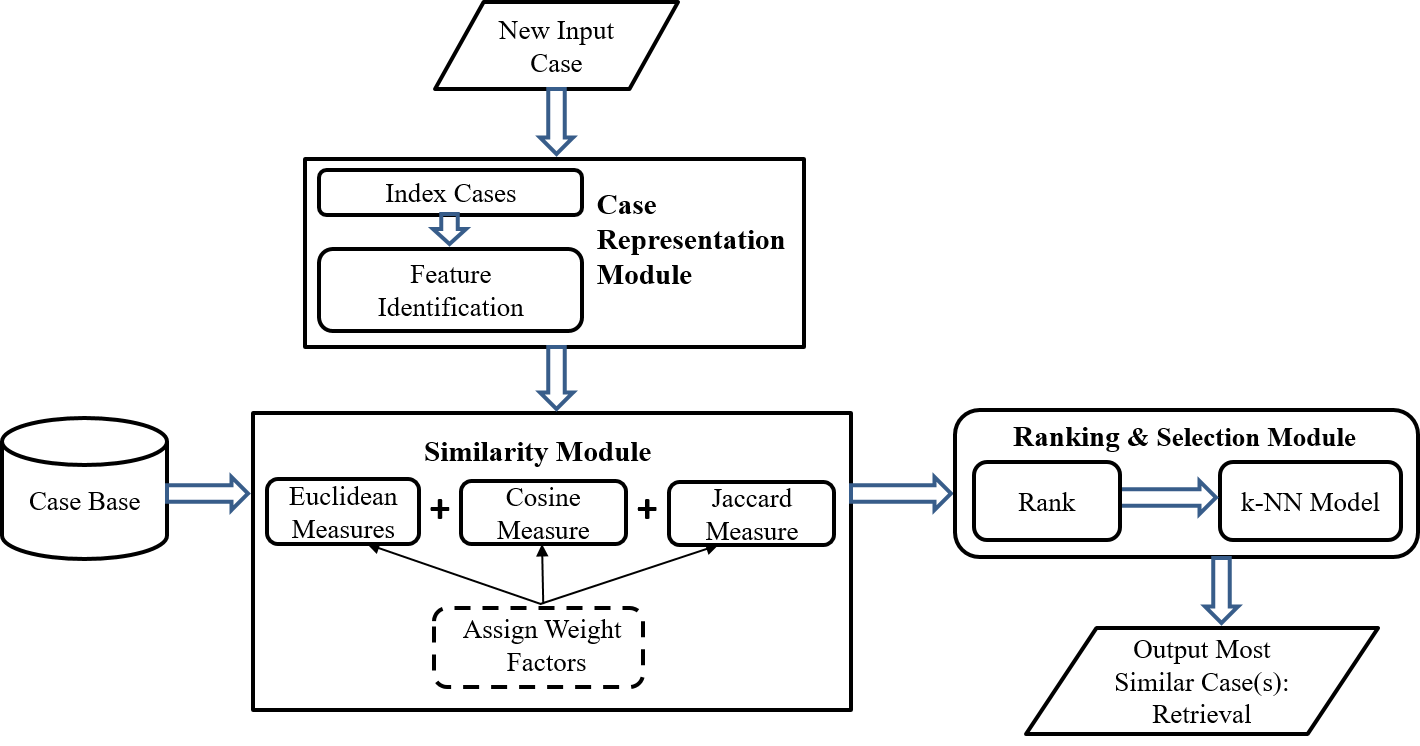
**Case Retrieval Algorithm**: The k-NN model was used in conjunction with the hybridized geometric similarity measure to identify retrieved cases whose value are higher the set threshold.

The functionality of the system include:

**Case Input**: The user provides information about the case to be tested.

**Case Retrieval**: The case retrieval algorithm employs the hybridized geometric similarity measure to identify the most similar cases from the case base.

* + - * 1. **Case Output**: Here, the most similar past case is extracted or identified. The case may need to undergo some form of adaptation or revision process if it is not a perfect solution for the new problem.

 Fig 3.2 Conceptual Diagram for the Case Retrieval (Source: Researcher, Omonijo O, 2024)

From the conceptual diagram in Fig 3. 2, already indexed cases whose features have been properly represented are stored in the case base in the form of problem description along with the corresponding solution. Such that, a case is stored as;

Case = {*Problem Description*, *Solution*}

Where *Problem Description* refer to a set of features which describes the problem and *Solution* represent a set of features which describe the solution. **Determining when a case should be considered for retrieval is the goal of indexing2. Selecting the appropriate features is very important as this will aid efficiency in retrieval.**

When a new case is presented to the model, the features of the new case (which are the attributes) are identified and compared with the features of already existing cases in the case base. The description of the features of the new problem forms a list of attributes and their corresponding value pair but with no solution features. **A description of a new case is given in Fig 3.3.**

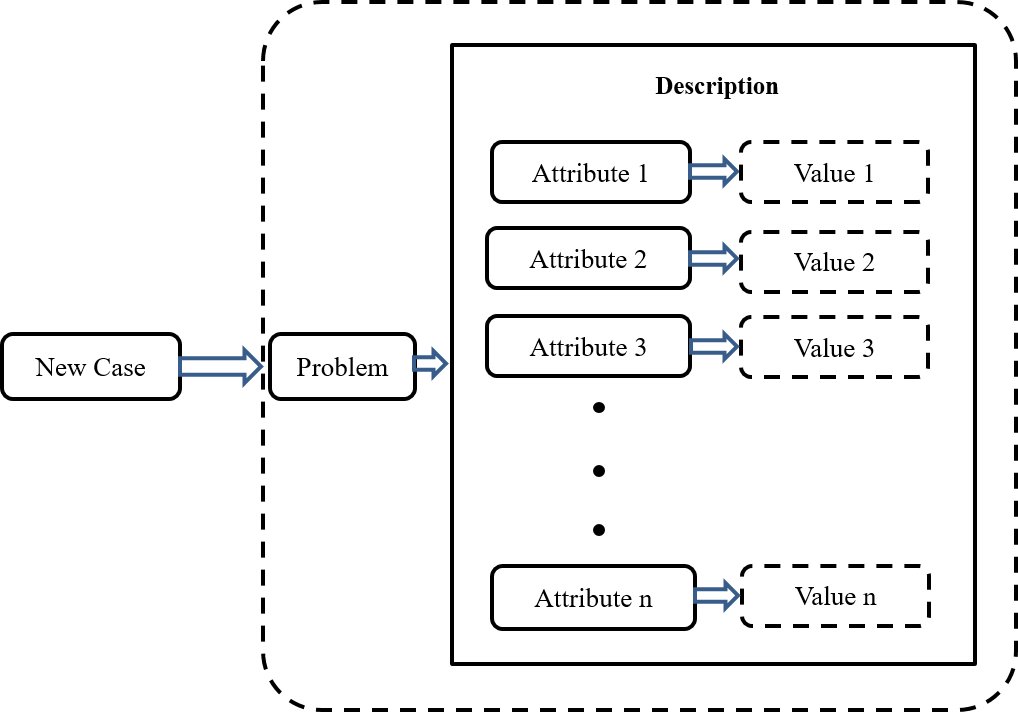


Fig 3. 3 Structure of a New Case (Source: Researcher, Omonijo O. 2024)

For each case in the case base, **weighted hybridized similarity scores** were calculated by using the proposed geometric similarity measure. Each of the similarity measure that make up the hybridized similarity score is assigned weights so as to signify the relative importance or contribution of each of them to the overall measurement. This is to ensure that there is a balance amongst the three measures such that the effect of the magnitude distance (Euclidean distance), orientation (Cosine similarity) and commonalities (Jaccard similarity) between new case and base cases will be considered. The assigned weights help to ensure that the most relevant similarity measure is given more weight in the similarity calculation, hence, leading to more accurate case retrieval. For example, if a measure is considered to be very important, it will be given a high weight so that it has a large impact on the similarity calculation. Conversely, if a measure is considered to be less important, it will be given a low weight so that it has a smaller impact on the similarity calculation. The total sum of the values of the three weights must be equal to one (1). This is to ensure that the value of the hybridized similarity measure does not exceed the normalized values of 0 and 1.

Thereafter, the cases are sorted in descending order based on their weighted hybridized similarity scores. Cases whose weighted hybridized similarity scores exceeds the set threshold are considered for selection. This implies that where *k* = *n*, the list containing all of the cases in memory or a small number of *k* whose value is greater than the similarity degree (threshold) are considered as potential solutions. Since the ranking is done in a decreasing order then, the top *k* cases which exceeds the threshold are selected from the ranked list to form the k-Nearest Neighbours. The k-NN model was used because it considers the similarity between the new problem and all of the cases in the case base. This gives a more global view of the similarity between the new problem and the case base hence, leads to retrieval of more accurate results.

If there exist an instance where no case exceeds the threshold, then the topmost cases would be considered by an expert and appropriate solution adapted and saved in the case base in case of future episodes.

The weights assigned to each measure can be adjusted based on the domain and data characteristics to prioritize different aspects of the composite similarity measures.

With respect to the proposed framework as shown in Fig 3.2, the similarity retrieval algorithm is defined using “local” and “global” similarity functions. The local similarity function measures the distance between the simple attributes by matching the features of the problem description attribute-by-attribute, whereas the global similarity function applies the results from the local similarity measures to compare the compound attributes. This is achieved by using the proposed hybrid geometric similarity measure. The simple attributes are the important features of each case while the compound attributes are the common features amongst aggregated cases. Finally, retrieved cases is based on similarity scores that exceed a set threshold.

A hybridized geometric similarity metric for case retrieval in CBR systems is developed using a combination of Cosine, Euclidian and Jaccard similarity measures which is used to build an improved case retrieval model for CBR systems based on k-NN algorithm and the evaluation of the performance of the proposed model with other traditional models is tested on a domain of diseases of catfish (Clarias Gariepinus).

### Algorithm for the Proposed System

The algorithm for the retrieval of similar past case by using the proposed hybrid geometric similarity measure is as follows:

**Step 0**: Initialize process.

**Step 1**: Enter problem description features of a new case which have been indexed and properly represented.

**Step 2**: Create empty lists of variables to store the k-NN and hybridized similarity scores.

**Step 3**: Set a threshold value.

**Step 4**: Loop through each case in the case base.

**Step 5**: For each case in the case base, measure their similarity with the new case by calculating the hybrid similarity scores and ensuring that the weights of each individual similarity measure is considered.

**Step 6**: Append the calculated similarity scores to the list created earlier.

**Step 7**: Sort the cases in descending order based on their hybrid similarity scores.

**Step 8**: Select the top *k* cases with the highest hybrid similarity scores which exceeds the set threshold as the k-Nearest Neighbours of the query case.

**Step 9**: Adapt the top-ranked cases to the specific context of the new problem, considering unique symptoms of the catfish disease.

**Step 10**: Output similar solutions (i.e. after the match, a list of similar solution to the new case will be displayed). Reuse the most similar case (one with the highest similarity value) as recommended solution to the new case.

**Step 11**: End process

A pseudocode of the retrieval process is as follow:

*<Case Retrieval Module>*

*Begin the process*

*// Initialize variables*

*Input the features of the problem description of the new case*

*k\_NN = [ ]*

*similarity\_scores = [ ]*

*threshold\_value*

*// Calculate the hybrid similarity scores for each case in the case base*

*for each case in case\_base:*

*for each attribute(q):value(q) pair of the new case associated with attribute(i):value(i) pair of cases in the case\_base:*

*for each case in case\_base:*

*individual\_scores = [ ]*

*// Calculate similarity using the measure, incorporating appropriate weights*

*hybrid\_individual\_score = sum(similarity\_measure, weights(query\_case, case, ))/3*

*similarity\_scores.append(hybrid\_similarity\_score)*

*// Rank similarity\_scores in descending order*

*// Find k-NN based on hybrid scores*

*k\_NN = top\_k(case\_base, similarity\_scores, k)*

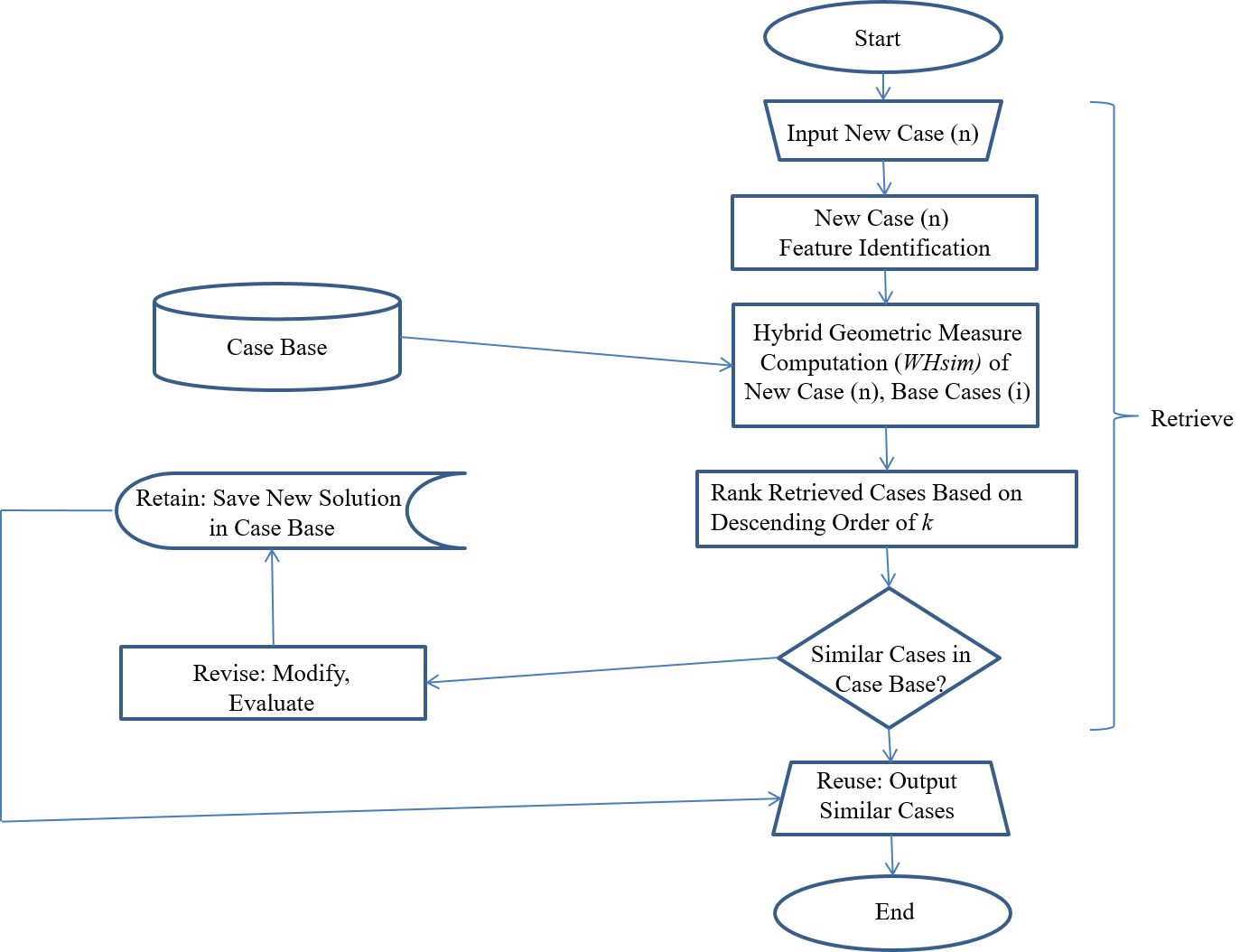
*return k\_NN based on the threshod\_value*

*Adapt the k-NN to fit new problem context*

*Output similar solutions*

*<End the process>*

From the algorithm presented, if given a new problem, the problem description (features) of the new case are extracted and represented in terms of attribute-value pairs. Then, the attribute-value pair of the new case are compared with each attribute-value pair of each of the case in the case base using the proposed hybrid geometric similarity measure. The cases are ranked and sorted in descending order based on their hybrid similarity score. The top *k* cases that exceeds the threshold value from the sorted list are selected as the k-Nearest Neighbours of the query case. The top-ranked cases are adapted to the specific context of the new problem. Finally, the most similar solution is output (i.e. after the match, a list of similar solution to the new case will be displayed) and reused as the most similar case (one with the highest similarity value) with recommended solution to the new case.

 Fig 3.4 Flowchart for the Hybrid Similarity Measure (Researcher, Omonijo O., 2024)

The flowchart in Fig 3.4 presents a breakdown of the flow of each stage of the proposed hybrid geometric similarity retrieval using k-NN.

To verify the effectiveness and performance of the proposed measure, the model was applied to a domain of catfish (Clarias Gariepinus) diseases and new cases are used to test the model. The level of success of the model is evaluated base on accuracy, precision, recall, F1-score, area under the ROC curve and robustness.

### Data Acquisition

CBR systems are data driven, hence the need for data to implement the proposed system. Data on the diseases of catfish was employed for this study. This is because it can be represented using a feature-value concept**.**

Fish play an important role in nutrition, food security, and livelihoods4. In fact, the human consumption of fish meat accounts for approximately 40% of protein consumed5.

There are different types of fishes and they basically inhabit in different water habitats. Catfish (Clarias Gariepinus) are primarily freshwater fish that are well adapted to confined environments and are resistant to manipulation and disease10.While fish serves as a source for the provision of very high quantity of protein and a wide range of other nutrients especially essential amino acids and fatty acids that our body needs, it is also associated with a variety of diseases. These diseases can be detected by their associated symptoms. Diseases of fishes are provoked by a spacious range of contagious organisms such as bacteria, viruses, protozoan, and metazoan parasites6. The symptoms associated with each disease can sometimes be misleading, thus, both primary and secondary methods of data collection were used to gather the essential domain knowledge in order to have a better understanding of the area under examination. Data from relevant materials from sources such as, the internet, books, journal articles and prior research on catfish diseases cases were consulted. Despite the fact that information may be created from a variety of sources such as textbooks, manuals, and simulation models, the knowledge at the core of a well-built knowledge-based system originates from human specialists7. Thus, interviews with relevant catfish experts were conducted to obtain data depending on their degree of expertise and availability. This helped to provide a better understanding symptoms and diseases of the domain knowledge. Human specialists frequently apply their expertise and experience to address real-world challenges such as this. And in such instances, professionals employ not only their heuristics knowledge, but also their experience, which necessitates subjective judgment in making decisions about the topic at hand. This further demonstrates how crucial it is to capture expert knowledge and the techniques by which they do their tasks when designing a knowledge-based system7. At this stage, it is important to take care and critically examine the sort of knowledge to be obtained while seeking to capture and maintain knowledge.

Altogether, a total of 32 catfish diseases was used and 35 symptoms. Example of the data used in the model is presented in Table 3.2. The specified data are taken from all of the aforementioned media.

Table 3. 2 Description of Common African Catfish Diseases and Symptoms

|  |  |  |
| --- | --- | --- |
| **S/N** | **DISEASE** | **SYMPTOMS** |
| **A** | **BACTERIAL DISEASES** | |
| **1** | Motile Aeromonad Septicemia (MAS) | Hemorrhages on the skin, fins, and internal organs |
| Abdominal swelling |
| Abscesses (internal and external) |
| Fin rot/fry |
| Exophthalmos (pop-eye) |
| Ulcers on the body |
| Lethargy |
| loss of appetite |
| Generalized septicemia (blood infection) |
| Erratic swimming |
| Darkening of the body |
| **2** | Pseudomonas Infection | Hemorrhagic septicemia |
| Skin lesions and ulcers |
| Fin rot and tail rot |
| Swelling of the abdomen/belly (dropsy, ascites) |
| Exophthalmos (pop-eye) |
| Respiration is difficult |
| Loss of Appetite |
| Exhibits an erratic swimming behavior. |
| Lethargy |
| Fluid accumulation under the skin |
| **3** | Enteric Septicemia of Catfish (ESC) | Hemorrhages on the skin and internal organs |
| Loss of appetite |
| Pale gills due to anemia |
| Swollen abdomen (ascites) |
| Multiple small white spots on the skin |
| Ulcers on the skin and muscle tissue |
| Protruding eyes (exophthalmos) |
| Erratic swimming behavior |
| Lethargy |
| **4** | Dropsy | Swollen abdomen due to fluid accumulation (ascites) |
| Lethargy |
| Fins clamped against the body |
| Increased respiratory rate/ |
| Stay at the surface of the water |
| Swollen anus |
| Curved spine |
| Loss of appetite |
| Pale or swollen gills |
| Redness or inflammation of the skin (discoloration) |
| Pop-eye |
| **5** | Vibriosis | Skin ulcers |
| Skin lesions |
| Pop-eye |
| Loss of appetite |
| Swollen abdomen (ascites) |
| Lethargy |
|  |
| **6** | Streptococcosis | Exophthalmos (pop-eye) |
| Hemorrhages on the skin, eyes, and fins |
| Skin ulcers and abscesses |
| Erratic swimming (loss of buoyancy control) |
| Loss of equilibrium |
| Lethargy |
| Loss of appetite |
| Corneal opacity (whitish eyes) |
| Ascites (i.e., distended abdomen/dropsy/bloating) |
| **7** | Epitheliocystis | White cysts or nodules on the gills or skin |
| Difficulty breathing due to gill damage |
| Lethargy |
| Increased mucus production on the gills |
| Fish swimming near surface |
| **8** | Bacterial Gill Disease (BGD) | Frayed and eroded gill filaments |
| Excessive mucus production on the gills |
| Labored breathing or gasping at the water surface (difficulty breathing) |
| Lethargy |
| Pale gills due to necrosis |
| Loss of appetite |
| Rapid gill movement |
| **9** | Columnaris Disease | Grayish-white spots on the skin, gills, and fins |
| Skin ulcers |
| Lesions on the skin |
| Frayed (eroded) fins |
| Mucus buildup on the gills |
| Difficulty breathing and increased respiratory rate |
| Lethargy |
| Loss of appetite |
| **B** | **VIRAL DISEASES** | |
| **10** | Ictalurid Herpesvirus 1 (IcHV-1): Channel Catfish Virus Disease (CCVD) | Swollen abdomen (ascites) |
| Lethargy |
| Exophthalmos (pop-eye) |
| Hemorrhages in the skin, fins, and eyes |
| Pale gills due to anemia |
| Erratic swimming behavior or spiraling |
| Darkened skin color |
| Enlarged spleen |
| Loss of appetite |
| **11** | Viral Hemorrhagic Septicemia (VHS) | Hemorrhages in the skin, gills, and eyes |
| Bulging eyes (Exophthalmos) |
| Anemia and pale gills |
| Swollen and darkened spleen |
| Lethargy |
| Loss of appetite |
| Erratic swimming and loss of equilibrium |
| External and internal bleeding |
| Swollen abdomen |
| **12** | Nervous Necrosis Virus (NNV) | Erratic swimming |
| Spiraling |
| Loss of appetite |
| Darkened skin |
| Lesions |
| Loss of balance (Buoyancy) |
| Muscle twitching |
| Pop-eye |
| **13** | Epizootic hematopoietic necrosis virus (EHNV) | Hemorrhages in the internal organs and muscles |
| Swollen abdomen |
| Darkening of the skin |
| Enlarged spleen and kidney |
| Loss of appetite |
| Focal white to yellow liver lesions |
| **14** | Viral Encephalopathy and Retinopathy (VER) | Erratic swimming |
| Loss of equilibrium Hyperactivity |
| Sporadic protrusion of the head from the water. |
| Colour change |
| Lethargy |
| Reduced feeding |
| Retinal lesions and eye abnormalities (Loss of vision) |
| Emaciation |
| Over-inflated swim bladder (the only significant internal gross pathological sign). |
|  |
| **C** | **PARASITIC DISEASES** | |
| **15** | Ichthyophthiriasis (Ich) | White spots on the skin, fins, and gills (resembling grains of salt or sugar) |
| Rubbing or scratching against objects (Flashing) |
| Discoloration |
| Upside-down swimming near the surface |
| Increased mucus production |
| Labored breathing (hyperventilation) |
| Rapid gill movement |
| Lethargy |
| Loss of appetite (Anorexia) |
| **16** | Trichinosis | Excessive mucus production on the gills and skin |
| Clamped fins |
| Rubbing or flashing against objects |
| Labored breathing and rapid gill movement |
| Lethargy |
| Skin discoloration |
| Ulcer of the skin |
| Reduced feeding |
| Frayed fins |
| Skin erosion |
| **17** | Dactylogyrosis (Gill Flukes) | Excessive mucus production on the gills and skin |
| Frayed or eroded gill filaments |
| Labored breathing and gasping at the surface |
| Lethargy |
| Flashing |
| Pale gills due to anemia |
| **18** | Gyrodactylosis | Excessive mucus production on the skin and gills |
| Rubbing or scratching against objects |
| Frayed fins |
| Skin damage |
| Pale gills |
| Labored breathing and rapid gill movement |
| Lethargy |
| Loss of appetite |
| Visible parasites on the skin and fins (using magnification) |
| **19** | Argulusosis | Visible parasites (flat, oval-shaped) attached to the skin |
| Rubbing or scratching against objects |
| Skin damage (redness and inflammation) at attachment sites |
| Lethargy |
| Loss of appetite |
| Excess mucus production |
| Secondary bacterial infections at attachment sites |
| **20** | Lerneaosis **(**Anchor worms**)** | Long, thin worms protruding from the body surface or gills |
| Skin ulcers at the point of attachment |
| Rubbing or scratching against objects |
| Lethargy |
| Reduced feeding |
| Lesions |
| Secondary bacterial infections |
| **21** | Proliferative Gill Disease (PGD) | Swollen and mottled gills |
| Excessive mucus production on the gills |
| Labored breathing and gasping at the surface |
| Lethargy |
| Reduced feeding |
| Pale gills due to necrosis |
| High mortality, especially in young fish |
| **22** | Flukes (Trematodes) | Excessive mucus production on the skin and gills |
| Rubbing or scratching against objects |
| Frayed fins |
| Skin damage |
| Labored breathing and rapid gill movement |
| Lethargy |
| Loss of appetite |
| Visible flukes under magnification |
| **23** | Epistylis | White to grayish cotton-like growths on the skin, fins, and gills (White or grayish cysts on the skin or gills) |
| Rubbing or scratching against objects |
| Increased mucus production |
| Labored breathing and gasping at the surface |
| Lethargy |
| Reduced feeding |
| Secondary bacterial infections due to damaged tissue |
| **24** | Myxomycosis | Cysts or spores in the muscles, skin, or internal organs (Ulcer) |
| Swollen body areas |
| Exophthalmos (pop-eye) |
| Difficulty swimming and erratic movements |
| Emaciation |
| Weight loss (Difficulty eating or drinking) |
| Lethargy |
| **D** | **FUNGAL INFECTIONS** | |
| **25** | Branchiomycosis  (More of internal signs) | Discoloration and necrosis of gill tissues, often with a mottled appearance (red and white patches) |
| Labored breathing and gasping at the water surface |
| Rapid gill movement |
| Lethargy |
| Reduced feeding |
| Pale gills due to impaired blood flow |
| Increased mucus production on gills |
| Mottled, or blotchy in appearance due to the dying tissue |
| **26** | Dermatomycosis | Grayish-white to brownish lesions on the skin, fins, and sometimes gills |
| Cotton-like fungal growths on affected areas (Fungal growth appears on the body surface (skin and fins)) |
| Ulcerations |
| Erosion of the skin |
| Loss of appetite |
| Lesion |
| **27** | Saprolegniasis (Winter Fungus) | White or grayish cotton-like growth on the skin, fins, or gills |
| Lesions |
| Skin ulcerations |
| Lethargy |
| depigmented skin |
| Sunken eyes (Endophthalmia) |
| Hemorrhagic margins/signs |
| **E** | **ENVIRONMENTAL STRESS-RELATED DISEASES AND NUTRITIONAL DEFICIENCIES DISEASES** | |
| **28** | Fungal Gill Disease | White or grayish fungal patches on the gills |
| Labored breathing |
| Increased gill movement |
| Gills may appear swollen or frayed |
| Lethargy |
| Look for excessive mucus production on the gills |
| Reduced feeding |
| Pale gills due to impaired oxygen exchange |
| **29** | Costiasis | Increased mucus production |
| Rubbing or flashing against objects |
| Labored breathing and rapid gill movement |
| Lethargy |
| Reduced feeding |
| Frayed fins |
| Skin erosion |
| Clamped fins |
| **30** | Gas Bubble Disease | Visible gas bubbles under the skin, in the eyes, or gills (appears like fish glitter) |
| Buoyancy problems (Difficulty staying afloat) |
| Gasping at the water surface |
| Lethargy |
| Sudden death in severe cases |
| **31** | Ammonia Toxicity | Red or inflamed gills |
| Gasping at the water surface |
| Lethargy |
| Reduced appetite |
| Rapid gill movement |
| Erratic swimming behavior |
| Skin or fin lesions |
| Sudden death in severe cases |
| **32** | Nitrite Toxicity | Brown or tan colored gills (brown blood disease) |
| Rapid gill movement |
| Labored breathing |
| Lethargy |
| Reduced feeding |
| Clamped fins |
| Sudden death in severe cases |

Table 3. 3 Common African Catfish Diseases List

|  |  |
| --- | --- |
| **Disease Code** | **Disease Name** |
| **Bacterial Diseases** | |
| **D1** | Motile Aeromonad Septicemia (MAS) |
| **D2** | Pseudomonas Infection |
| **D3** | Enteric Septicemia of Catfish (ESC) |
| **D4** | Dropsy |
| **D5** | Vibriosis |
| **D6** | Streptococcosis |
| **D7** | Epitheliocystis |
| **D8** | Bacterial Gill Disease (BGD) |
| **D9** | Columnaris Disease |
| **Viral Diseases** | |
| **D10** | Ictalurid Herpesvirus 1 (IcHV-1): Channel Catfish Virus Disease (CCVD) |
| **D11** | Viral Hemorrhagic Septicemia (VHS) |
| **D12** | Nervous Necrosis Virus (NNV) |
| **D13** | Epizootic hematopoietic necrosis virus (EHNV) |
| **D14** | Viral Encephalopathy and Retinopathy (VER) |
| **Parasitic Diseases** | |
| **D15** | Ichthyophthiriasis (Ich) |
| **D16** | Trichodinosis |
| **D17** | Dactylogyrosis (Gill Flukes) |
| **D18** | Gyrodactylosis |
| **D19** | Argulusosis |
| **D20** | Lerneaosis **(Anchor worms)** |
| **D21** | Proliferative Gill Disease (PGD): Hamburger Gill Disease |
| **D22** | Trematodes |
| **D23** | Epistylis |
| **D24** | Myxomycosis |
| **Fungal Infections** | |
| **D25** | Branchiomycosis |
| **D26** | Dermatomycosis |
| **D27** | Saprolegniasis **(Winter Fungus)** |
| **Environmental Stress-Related Diseases and Nutritional Deficiencies Diseases** | |
| **D28** | Fungal Gill Disease |
| **D29** | Costiasis |
| **D30** | Gas Bubble Disease |
| **D31** | Ammonia Toxicity |
| **D32** | Nitrite Toxicity |

Table 3. 4 Feature List of Symptoms

|  |  |  |
| --- | --- | --- |
| **SN** | **Feature Code** | **Symptoms/Features Description** |
| **1** | S1 |  |
| **2** | S2 |  |
| **3** | S3 |  |
| **4** | S4 |  |
| **5** | S5 |  |
| **6** | S6 |  |
| **7** | S7 |  |

After acquisition of the data, the data needs to be represented in the form that is acceptable for storage in the system’s case base for the purpose of retrieval process. The case structure was created to convey knowledge in a feature-value pair style. Where the problem description and solution are the two most significant aspects of the case structure. Problem description attributes such as the fish's Case ID, Case Feature and Case Alias were considered. As shown in Table 3.1, the problem description is made up of characteristics (Case ID, Case Feature and Case Alias) that specified the difficulties to be handled as part of the case structure. The hybridized geometric similarity measure then considers these factors to identify similar cases and potentially provide an accurate disease type. Based on the problem descriptions, the solution section is suggested. The identification of infected fish timely is an obligatory step to thwart from spreading disease6.

### System Architecture

The system architecture consists of the following components:

**Case Base**: The case base is a repository of cases that have been solved in the past. The following fields make up the case base: Case ID, Case Feature, Case Alias (optional) and Solution.

**Hybridized Geometric Similarity Measure**: The hybridized geometric similarity measure is used to compute the similarity between cases. The similarity measure is based on a combination of existing geometric similarity measures, such as the Euclidean distance measure, cosine similarity measure and Jaccard similarity measure.

**Case Retrieval Algorithm**: The case retrieval algorithm uses the hybridized geometric similarity measure to retrieve the most similar cases from the case base for a new problem.

### 3.3.3 Technologies Used

The following technologies were used to implement the system:

**Programming Language**: Python

**Database/Case Library**: MySQL

**Case Base Management System**: Apache Mahout

**Hybridized Geometric Similarity Measure Library**: NumPy or SciPy

### 3.3.4 Materials Used

The materials used to conduct the study on improved case retrieval in CBR systems with a hybridized geometric similarity measure are:

* + - * 1. **Dataset of Cases**: A dataset of cases in the domain of fish diseases.
        2. **Software**: The programming language, database, case base management system, and hybridized geometric similarity measure library listed in subsection 3.3.3.

## Requirements Specification

For the proposed CBR retrieval model to function efficiently, certain requirements were put in place. This requirements specification outlines the design and functionality of the CBR system employing the proposed hybridized geometric similarity measure and k-NN model. By meeting these requirements, the system aims to provide an accurate, efficient case retrieval and user-friendly CBR system. The system can be customized and extended to cater to specific needs and datasets, making it a versatile tool for problem solving and decision making. The subsections in this section discusses the requirements specification needed for the CBR system which utilizes the proposed hybridized geometric similarity measure with k-NN to function efficiently.

### System Requirements

This requirements specification outlines the design and functionality of a fish disease detection system utilizing a CBR approach with k-NN integration. The system aims to improve fish health management by providing accurate diagnoses.

#### Functional Requirements

The following are the functional requirements needed for efficient operation of the model:

1. **Case Input**: Users are required to input information about a new case. With respect to the data set used in the study, the system allows users to input information about new fish disease cases, including: symptoms and disease alias.
2. **Case Base**: The system has a well-structured and annotated case base containing fish disease cases with detailed information on fish disease signs, disease details. Each case in the case base has a case identification (ID). This support efficient indexing and retrieval of cases based on user queries. The case base is easily maintainable and updatable to incorporate new knowledge and improve the system's performance.
3. **Case Retrieval**: The system employs the proposed hybrid similarity measure which combines traditional geometric similarity measures (Euclidean, Cosine, Jaccard) with k-NN to retrieve the most relevant cases from the case base based on the new fish disease case presented by the user. The system calculates individual similarity scores using hybridized geometric measures and prioritize retrieved cases based on their global similarity score obtained through k-NN. That is, rank retrieved cases based on their global similarity scores to the problem at hand.
4. **Output and User Interface**: The system presents solution in a clear and concise manner and provide visualization tools for exploring case similarities and relationships. This allows user interaction to refine queries, provide additional information, and receive clarifications.

#### Non-Functional Requirements

**Performance**: The system responds to user queries promptly and efficiently. The system is accurate, with a high precision and recall rate in identifying relevant cases and providing accurate diagnoses.

**Accuracy**: The hybridized similarity measure and k-NN is utilized to achieve high accuracy in case retrieval and diagnosis/treatment recommendations. It provides confidence levels (ranking score) for recommendations to indicate the uncertainty involved.

**Usability**: The system's interface is user-friendly and intuitive, allowing users with varying levels of technical expertise to easily input data and access results.

Scalability: The system is scalable to accommodate larger case bases and increasing user demand. But then, the significant performance degradation has not been ascertained.

### 3.4.2 Programming Techniques

The following are the programming technologies used to implement the system:

**Programming Language**: Python programming language was used to implement the system due to its extensive libraries for mathematics, machine learning, and data analysis.

**Machine Learning Libraries and Frameworks**: Machine learning libraries (such as NumPy, pandas and scikit-learn) were used for numerical computation and other data analysis tasks, case representation data structures, similarity calculations and graphical user interface (GUI) libraries for user interaction.

**Database/Case Library**: A suitable Relational Database Management System (RDBMS) was used for storing and managing the case base. The RDBMS that was used for storing and querying of the case base in the system is MySQL.

### System Requirements

**Minimum Hardware**:

Personal Computer with a modern processor (e.g., Intel Core i5 or AMD Ryzen 5 or equivalent)

Sufficient RAM (e.g., 8GB)

Storage: 500GB HDD or SSD.

**Operating System**: Operating system such as Windows 10/11, mac OS, Linux.

Web browser for accessing the system interface with JavaScript enabled.

### Input Processing

The system accept user input through a user-friendly interface to describe the new problem or issue at hand.

The input data are validated for accuracy and completeness before processing to ensure quality and consistency before being used for case retrieval. The system handles missing data gracefully and provide appropriate warnings to the user. Data normalization techniques are applied to ensure compatibility with similarity measures.

### Output Processing

The output is such that it generates textual reports with graphical representations of case similarities and relationships.

### Testing and Validation

The system is thoroughly tested on a representative dataset of fish disease cases. Accuracy metrics is used to evaluate the effectiveness and efficiency of the proposed model. User testing was conducted to assess the usability and effectiveness of the system interface.

## Integrated Development Environment

In the implementation of this model, the Jupyter Notebook Independent Development Environment (IDE) was used. Jupyter Notebook IDE was chosen because it is excellent for interactive coding and data exploration. Also, it is well-suited for experimenting with algorithms and visualizing data. Hence, it is well-suited for research of this nature.

## Evaluation Metrics

The performance of the proposed case similarity retrieval model is evaluated on the developed fish diseases dataset. The evaluation compares the performance of the CBR system to other CBR systems that use traditional similarity measures for case retrieval. Evaluation metrics that were used to assess the performance of the hybridized geometric similarity measures and the CBR system are discussed in the forgoing subsections.

### Accuracy

Measures the overall success rate of the system in correctly identifying relevant cases. Higher accuracy indicates better retrieval performance.

**Accuracy** =

### Precision

Precision measures the proportion of correctly identified positive cases (true diagnoses) among all cases predicted as positive. Measures the proportion of retrieved cases that are actually relevant. Higher precision indicates fewer irrelevant cases retrieved.

**Precision** =

### Recall

Measures the proportion of actual positive cases that are correctly identified. Recall measures the proportion of true diagnoses made among all actual positive cases. Higher recall indicates fewer relevant cases missed.

**Recall** =

### F1-Score

F1-Score combines precision and recall into a single metric, providing a balanced view of the system's performance.

**F1-Score** = 2 \*

### Area Under the ROC Curve (AUC)

Independent of class distribution, useful when dealing with imbalanced datasets. Measures the overall ability of the system to discriminate between relevant and irrelevant cases. It is calculated based on True Positive Rate and False Positive Rate.

### Case-Based Metrics

The Nearest Neighbour Average Precision (NNAP) is used to measure the average precision of the retrieved nearest neighbours. Measures the proportion of cases in the test set that have at least one relevant case retrieved.

Case retrieval system development: Develop a CBR system that uses one of the new hybridized geometric similarity measures for case retrieval.